DOCUMENT RESUME

ED 052 737 HE 002 368

AUTHOR Hillery, Joseph M.; Yukl, Gary A.
TITLE Convergent and Discriminant Validation of Student Ratings of College Instructors.
INSTITUTION Akron Univ., Ohio.
PUB DATE 8 May 71
NOTE 13p.; Paper presented at the Midwestern Psychological Association Convention, Detroit, Michigan, May 8, 1971

EDRS PRICE EDRS Price MF-$0.65 HC-$3.29
DESCRIPTORS *Effective Teaching, *Higher Education, *Student Opinion, Teacher Evaluation, *Teacher Rating, *Validity
IDENTIFIERS *Akron University

ABSTRACT

This paper reports the results of a validation study of data obtained from a teacher rating survey conducted by the University of Akron Student Council during the Fall 1969. The rating questionnaire consisted of 14 times: two items measured the student's overall evaluation of his instructor; 5 items measured specific performance dimensions such as stimulation, communication, consideration, evaluation, and workload, and each of these dimensions was measured by two methods: (1) asking the student to compare his instructor with others he had known, and (2) requiring the student to make an absolute evaluation of the instructor on a graphic rating scale. The last two items obtained information on the student's class standing, and his cumulative GPA. Information was also obtained on the size of each class, the average grade given in each course, and the instructor's rank. The data analysis consisted of the multitrait, multimethod approach to convergent and discriminant validation, first proposed by Campbell and Fiske in 1959. The results indicated that the performance dimensions showed fairly high reliability and convergent validity. However, the discriminant validity was not high enough to conclude that independent dimensions of instructor performance were being accurately measured. (AF)

ED052737

Convergent and Discriminant Validation of Student Ratings of College Instructors

Joseph M. Hillery and Gary A. Yuki

The University of Akron

Paper Presented at the

Midwestern Psychological Association

Convention, Detroit, May 8, 1971

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

HE 002 368

Introduction

Although everyone agrees on the importance of good teaching, little is known about what makes a good teacher. It is not yet clear what aspects of an instructor's behavior are the most essential in achieving the dual goals of student learning and student satisfaction with their educational experience. Before this can be determined, it is necessary to identify separate dimensions of instructor performance and to develop accurate measures of these performance dimensions. This paper reports the results of a validation study of data obtained from a survey conducted by the University of Akron Student Council during the Fall quarter of 1969.

Method

The rating questionnaire consisted of 14 items and was distributed to students in their classrooms during the last week of class. Two items in the questionnaire (items 11 & 12) measured the students' overall evaluation of their instructor. In addition, five specific performance dimensions which seemed to be separate and meaningful were measured. These specific performance dimensions were labeled and defined as follows:

Stimulation: How well is the instructor able to stimulate student interest and enthusiasm in the course? (items 1A & 2A)

Communication: How clear and well-organized are the instructor's lectures or explanations? (items 1B & 2B)

Consideration: How friendly, helpful, approachable, and considerate is the instructor? (items 1C & 2C)

Evaluation: How objective, fair, and comprehensive is the instructor's grading of students? (items 1D & 2D)

Workload: How heavy and demanding is the course workload (e.g. reading, assignments, and requirements)? (items 1E & 2E)

These five dimensions were selected after reviewing the results of previous studies involving factor analysis of student ratings. Each dimension was measured by two methods. Method 1 called for a relative evaluation of the instructor; that is, the student was asked to compare his instructor with others he had known. Method 2 required the student to make an absolute evaluation of the instructor on a graphic rating scale.

Two additional items obtained information concerning the student's class standing (item 13) and his cumulative G.P.A. (item 14). From the university records, the following information was obtained: the size of each class, the average grade given in each course, and the instructors rank.

The data analysis consisted of the multi-trait, multi-method approach to convergent and discriminant validation, first proposed by Campbell and Fiske in 1959. As applied to the student rating

data, the procedure involved calculating intercorrelation matrices and examining the pattern of correlations to determine if the ratings from two different methods of measuring a single dimension agreed more than ratings on two different dimensions measured by a single method.

Results

The ratings given by the students showed considerable variety in their responses. This was demonstrated, not only in ratings given by the individuals, but for average class ratings as well. Taking one of the overall evaluation items for an example, when the mean rating for each class was calculated, the range of the class means was from 0.4 to 4.0, which is only slightly smaller than the total possible range of from 0 to 4.0. This distribution of class means was negatively skewed, showing a tendency for leniency in student ratings of their instructors. Although the midpoint of the scale was at 2.00, the actual mean of the average class ratings was 2.74.

The first analysis of convergent and discriminant validity is represented in Table 1. This table presents a correlation matrix indicating the correlation between each possible pair of items measuring the five specific performance dimensions. Each circled value is the correlation between the two types of methods (i.e. relative and absolute ratings) for a single dimension. The higher

the circled correlations, the better the convergent validity. In this analysis, convergent validity is similar to internal consistency reliability. Convergent validity for Stimulation, Communication, Consideration, and Evaluation was fairly high, but not as high as we would have hoped. However, there was an obvious lack of agreement between the two items measuring workload. This may have been due to a response set built up from the pattern of previous response alternatives. If you look at the preceding questions using Method 1, you will see that "considerably above average" and "above average" were response choices which represented a high evaluation of the instructor. Some students may have interpreted the response choices for the Workload item (item 1E) in the same way. However, for the Workload item, "considerably above average" and "above average" were supposed to indicate an above average workload, not an above average evaluation of the instructor.

Discriminant validity is evaluated by comparing a circled value with the other correlation values in the same row and column in the matrix. The lower these other values are, and the greater the difference between the cirled value and these other values, the better is discriminant validity. The discriminant validity of the specific performance dimensions was only mildly impressive. There are several possible explanations for the lack of clear-cut discriminant validity: (1) the dimensions are not really inde-

pendent, (2) the ratings are contaminated by the particular measuring procedure which is used, or (3) the raters are susceptible to a general evaluative halo. Although it is not possible to determine to what extent each explanation is correct, we believe that the halo effect is the most likely explanation for our results.

A second analysis of convergent and discriminant validity is based on Table 2. Each class was randomly divided into two equal groups of raters, and the extent of agreement between the two groups was determined for each of the items. For this analysis, both items measuring a dimension were combined. Convergent validity for Stimulation, Communication, Consideration, and Evaluation was very impressive, as evidenced by the very high circled values. However, Workload again showed low convergent validity. Since, in this analysis, the methods are actually randomly assigned groups of raters, convergent validity is somewhat similar to inter-rater reliability. Discriminant for this matrix was low, the circled values are not much larger than the other values in the same row or column, and these other values are large, which is not desirable. These results lead one to conclude that the graphic rating scale was not measuring separate and independent aspects of instructor performance.

Evidence concerning the reliability of the two overall ratings (items 11 & 12) was also available. The correlation between these two items was .75 indicating adequate internal consistency.

reliability. The correlation between the two groups of raters in each class was .91 for item 11 and .92 for item 12, indicating high inter-rater reliability.

The results of the check for contamination of the overall ratings by various other variables are presented in Table 3. There appears to be no appreciable relationship between the ratings made by a student and his Grade Point Average or class standing (i.e. Freshman, Sophomore, Junior, or Senior). Furthermore, the ratings do not seem to be affected by the size of the class or the rank of the instructor (i.e. instructor, assistant professor, associate professor, or full professor). We did find a correlation of .29 between the average grade given in the course and the average rating received by the instructor. However, this correlation accounts for only 9% of the total variance of the ratings and does not appear to be a serious contaminant.

Summary and Conclusions:

The performance dimensions measured by the student ratings showed fairly high reliability and convergent validity. However, the discriminant validity was not high enough to conclude that independent dimensions of instructor performance were being accurately measured. Since ratings on the specific performance dimensions were highly inter-correlated, it appears that all of the rating scale items were measuring the same dimension --probably

the student's general satisfaction with his instructor. Therefore, in order to simplify the administration and scoring procedures, it would be possible to use only the two general evaluation items and omit the specific performance items altogether. This shorter rating form would suffice as long as the general items were reliable and only a single overall rating is needed for each instructor. However, if additional information concerning specific traits or behaviors is desired, in order to provide diagnostic feedback to the instructor, then a method other than the graphic rating scale should be considered. A checklist or forced choice scale may prove more successful for this purpose.

In conclusion, the multi-trait, multi-method technique of estimating convergent and discriminant validity, does appear useful in evaluating student ratings. In our study the two rating methods were very similar to each other and this type of analysis would be even more meaningful if two very different rating methods were used, such as graphic ratings and forced choice. In any case the multi-trait, multi-method approach, yields a good deal of useful information about the reliability and validity of student ratings.

METHOD 1

Select the alternative which best describes how your instructor compares with instructors you have had in college. If you are a freshman please use your past college instructors and high school teachers for your comparison.

1A. How well is the instructor able to stimulate the student interest and enthusiasm in the course?

a) Considerably above average
b) Above average
c) Average
d) Below average
e) Considerably below average

1B. How clear and well organized are the instructor's lectures or explanations? Consider effectiveness of getting across the material to the student. (If no lecture please leave blank)

a) Considerably above average
b) Above average
c) Average
d) Below Average
e) Considerably below average

1C. How friendly, helpful, and considerate is your instructor?

a) Considerably above average
b) Above average
c) Average
d) Below average
e) Considerably below average

1D. How objective, fair, and comprehensive is the instructor's evaluation (i.e. grading) of your knowledge of the course material?

a) Considerably above average
b) Above average
c) Average
d) Below average
e) Considerably below average

1E. How heavy and demanding is the course workload, i.e. the reading and assignments? (If there is no reading or assignments please skip this question.)

a) Considerably above average
b) Above average
c) Average
d) Below average
e) Considerably below average

METHOD 2

The letters A to E in the following rating scales mark equal intervals along a continuum between two opposite extremes. For each scale rate your instructor by selecting the letter which best locates him with respect to the evaluative continuum.

2A. How interesting and stimulating is the instructor?

Very interesting A B C D E Very dull and boring

2B. How clear and well organized are the instructor's lectures or explanations? Consider effectiveness of getting across the material to the students.

Very clear and organized A B C D E Confusing and disorganized

2C. How friendly, helpful, and considerate is your instructor?

Very friendly and helpful A B C D E Hostile or inconsiderate

2D. How fair, objective, and comprehensive is the instructor's evaluation (i.e. grading) of your knowledge of course material?

Very fair and objective A B C D E Unfair and inadequate

2E. How difficult is the workload, that is, the assignments and reading? If none please leave blank.

Unusually easy workload A B C D E Very heavy workload

OVERALL RATINGS

11. In general, with A equal to the highest grade and F equal to the lowest grade, how would you rate this instructor's teaching of this course.

a) A
b) B
c) C
d) D
e) F

12. In general, how satisfied are you with your instructor?

Very satisfied A B C D E Very dissatisfied

Table 1

Multi-dimension, Multi-method Matrix for Two Methods and Five Dimensions of Instructor Performance*

		Method 1					Method 2				
	Dimension (item)	(1A)	(1B)	(1C)	(1D)	(1E)	(2A)	(2B)	(2C)	(2D)	(2E)
Method 1	Stimulation (1A)										
	Communication (1B)	.62									
	Consideration (1C)	.59	.45								
	Evaluation (1D)	.49	.43	.52							
	Workload (1E)	.21	.22	.24	.19						
Method 2	Stimulation (2A)	(.78)	.60	.57	.49	.21					
	Communication (2B)	.64	(.67)	.51	.48	.20	.72				
	Consideration (2C)	.53	.41	(.76)	.50	.22	.60	.56			
	Evaluation (2D)	.48	.41	.51	(.70)	.15	.53	.55	.59		
	Workload (2E)	.21	.21	.24	.27	(.00)	.25	.30	.30	.34	

*Note: N = 16,000 rating forms

Table 2

Multi-dimension, Multi-rater Matrix for Two Independent Groups of Raters and Five Performance Dimensions*

		Rater Group 1					Rater Group 2				
	Dimension (items)	(1A+2A)	(1B+2B)	(1C+2C)	(1D+2D)	(1E+2E)	(1A+2A)	(1B+2B)	(1C+2C)	(1D+2D)	(1E+2E)
Rater Group 1	Stimulation (1A+2A)										
	Communication (1B+2B)	.91									
	Consideration (1C+2C)	.75	.66								
	Evaluation (1D+2D)	.77	.74	.82							
	Workload (1E+2E)	.45	.46	.35	45						
Rater Group 2	Stimulation (1A+2A)	(.93)	.04	.71	.73	.40					
	Communication (1B+2B)	.88	(.92)	.63	.69	.42	.91				
	Consideration (1C+2C)	.71	.62	(.87)	.75	.32	.75	.68			
	Evaluation (1D+2D)	.71	.67	.72	(.86)	.40	.75	.72	.80		
	Workload (1E+2E)	.35	.33	.30	.38	(.40)	.36	.37	.36	.39	

*Note: N = 302 classes; all television courses and classes with less than 20 persons were omitted from this analysis.

Table 3

Correlation Between Overall Ratings

and Various Extraneous Variables

	Pearson r	Sample Size
Student Grade Point Average	.08	16,000[a]
Class Standing of Student	.13	16,000[a]
Class Size	-.07	435[b]
Instructor Rank	-.04	302[c]
Average Course Grade	.29	100[d]

[a]Note: This is the number of rating forms, not the number of students. Students usually rated more than one instructor during the survey.

[b]Note: All television courses were omitted from this computation.

[c]Note: All television courses and classes with less than 20 students were omitted from this computation.

[d]Note: These classes were selected randomly from non-television courses, with class sizes ranging from 20 to 100.